Does adaptive protein evolution proceed by large or small steps at the amino acid level?

Article (Accepted Version)

Does Adaptive Protein Evolution Proceed by Large or Small Steps at the Amino Acid Level?

Juraj Bergman[1,2]

Adam Eyre-Walker[3]

[1] Institut für Populationsgenetik, Vetmeduni Vienna, Veterinärplatz 1, Wien, A-1210, Austria

[2] Vienna Graduate School of Population Genetics, Wien, A-1210, Austria

[3] School of Life Sciences, University of Sussex, Brighton, BN1 9QG, United Kingdom

Abstract

A long-standing question in evolutionary biology is the relative contribution of large and small effect mutations to the adaptive process. We have investigated this question in proteins by estimating the rate of adaptive evolution between all pairs of amino acids separated by one mutational step using a McDonald-Kreitman type approach and genome-wide data from several *Drosophila* species. We find that the rate of adaptive evolution is highest amongst amino acids that are more similar. This is partly due to the fact that the proportion of mutations that are adaptive is higher amongst more similar amino acids. We also find that the rate of neutral evolution between amino acids is higher amongst more similar amino acids. Overall our results suggest that both the adaptive and non-adaptive evolution of proteins is dominated by substitutions between similar amino acids.

Introduction

Whether evolution proceeds by large or small steps is an old evolutionary problem that dates back, in its most extreme form, to the debate between saltationists and gradualists at the turn of the 20th century. It is a problem that is far from resolved despite extensive theoretical and experimental work (Barrett and Schluter 2008; Bell 2009; Rockman 2012).

There are in fact three related questions relating to the contribution of large and small mutations to the adaptive process: what is the distribution of effect sizes amongst new mutations, what is the distribution amongst those that spread to fixation, and is the process of adaptation largely a consequence of large or small mutations. An analogy might help to

illustrate the difference between the last two questions. Let us imagine that a builder is constructing a wall. The supply of bricks may be dominated by either large or small bricks and depending on her preferences for bricks, three different walls may be built; one in which most of the bricks are small and the wall is largely constructed of small bricks, one in which most of the bricks are small but the wall is largely built of large bricks and one in which most of the bricks are large and the wall is largely composed of large bricks.

Fisher (1930) originally suggested, based on his geometric model, that most advantageous mutations would be of small effect. While some experiments have been consistent with this expectation (Sanjuan, et al. 2004; Kassen and Bataillon 2006; Bataillon, et al. 2011; Schenk, et al. 2012) others have found a relatively uniform (Ferris, et al. 2007; MacLean and Buckling 2009) or normal distribution of effects (McDonald, et al. 2011). The difference between these studies seems to be largely a consequence of two factors; a tendency to under-sample mutations with weak effects because they are difficult to detect and how far the population is from the optimum. The further the population is from the optimum the more large-effect mutations are found (MacLean and Buckling 2009).

The distribution of mutant effects is however not the distribution of mutations fixed during evolution because large effect mutations have a greater chance of spreading to fixation (Kimura 1983). Theoretical work has suggested that the distribution of effects amongst mutations that spread to fixation is likely to be dominated by mutations of small effect if adaptation comes from new mutations, the underlying distribution of mutant effects is of the Gumbel (e.g. a normal distribution) or Weibull (e.g. a distribution with a truncated right tail) type and the fitness optimum moves suddenly (Orr 1998, 2002; Martin and Lenormand 2008) (though see critique by (Kopp and Hermisson 2009)). However, if the optimum moves slowly, or most adaptation comes from standing genetic variation (Barrett and Schluter 2008; Pritchard, et al. 2010), then substitutions of intermediate effect are expected to dominate the adaptive process (Kopp and Hermisson 2009; Matuszewski, et al. 2014, 2015). The distribution of substitution effects may be dominated by large effect mutations if the underlying distribution of mutant effects is heavy tailed (i.e. in the Frechet domain) (Seetharaman and Jain 2014).

Experiments that have tracked mutations that either fix or spread to high frequency under positive selection have found that the distribution can be dominated by mutations of small (Imhof and Schlotterer 2001; Perfeito, et al. 2007) or intermediate effect (Rozen, et al. 2002; Rokyta, et al. 2005; Barrett, et al. 2006; MacLean and Buckling 2009; Schoustra, et al. 2009). This again seems to depend on how far the population is from the optimum. If the population is far from the optimum, as in the experiment of Barrett et al. (Barrett, et al. 2006), then the distribution of mutations that rise to appreciable frequency, or are fixed, is dominated by intermediate or large effect mutations, because the distribution of new mutations is dominated by larger effect mutations (see above) and such mutations have a greater chance of spreading through the population. A second factor also comes into play in these experiments, which are usually conducted with asexual organisms – clonal interference. If there is clonal interference, then only mutations with intermediate or large effects can spread to high frequency or fixation (see (Perfeito, et al. 2007)).

These experiments have been very informative. However, all experiments assume that adaptation comes from new genetic variation, but this process might be dominated by standing genetic variation (Barrett and Schluter 2008; Pritchard, et al. 2010). Furthermore, clonal interference occurs in many of the experiments and it is not clear how many organisms are sufficiently asexual for this process to play an important role in adaptation. Finally, we have no idea whether evolution is dominated by large jumps in the optimum, as might be caused by the introduction of an antibiotic or a pesticide into the environment, or more gradual changes. The only experiments that would seem to give us information about what happens in the natural world are QTL analyses of the differences between species. These seem to suggest that much adaptation is due the fixation of mutations of large effect (Bell 2009), but as Rockman (Rockman 2012) has argued, some caution must be exercised because a single QTL may involve many mutations of smaller effect. Furthermore, such analyses only address the third question about adaptation; whether the adaptation is largely due to mutations of large or small effect.

Here we investigate whether adaptive evolution in proteins is dominated by mutations and substitutions between amino acid that are more or less similar to each other in their physicochemical properties. Grantham (Grantham 1974) and Miyata et al. (Miyata, et al.

1979) showed many years ago that the rate of amino acid substitution is negatively correlated to the difference in polarity, volume and chemical composition of the amino acids involved (see also (Zhang 2000)). This could be due to mutations between more different amino acids being either more deleterious or less advantageous. Gojobori et al. (Gojobori, et al. 2007) went a step further and showed that adaptive evolution was only detectable amongst amino acids that were more different. However, their analysis had several short-comings; in estimating the level of adaptive evolution they did not take into account slightly deleterious mutations, which bias estimates of adaptive evolution downwards; so, a lack of evidence of adaptive evolution between similar amino acids may simply be due to the segregation of slightly deleterious mutations. In our analysis, we estimate the rate of adaptive substitution between all pairs of amino acids separated by one mutational step using polymorphism data from *Drosophila melanogaster* polarized using *D. simulans* and *D. yakuba*, using a method which corrects for influence of slightly deleterious mutations on estimates of the rate of adaptive evolution. We also investigate whether mutations of large or small effect are more common and whether small or large steps contribute most to the increase in fitness.

Results

To investigate whether adaptive evolution is dominated by large or small steps at the molecular level we estimated the rate of adaptive evolution between all 75 pairs of amino acids that are separated by a single mutational step. We estimated the rates of substitution between *Drosophila melanogaster* and the *D. simulans/D. yakuba* outgroup pair using the method of Schneider et al. (Schneider, et al. 2011). This method is a variant of the McDonald-Kreitman (McDonald and Kreitman 1991) approach in which the rate of adaptive evolution is estimated by comparing the divergence at selected non-synonymous and neutral synonymous sites, to levels of polymorphisms at those same sites. The method estimates the distribution of fitness effects (DFE) of the neutral and deleterious non-synonymous mutations, assuming the DFE is described by a gamma distribution; the gamma distribution is characterised by the shape parameter, β, and the mean strength of selection acting against deleterious mutations multiplied by the effective population size, $N_e\bar{s}_d$. The method also estimates the proportion of mutations that are advantageous (λ_a) and the

strength of selection acting upon them multiplied by the effective population size ($N_e s_a$), as well as the rate of adaptive evolution relative to the mutation rate (ω_a) (Gossmann, et al. 2010). We initially focus our analysis on two properties of amino acids, volume and polarity, since these are two properties that all amino acids share and that have been studied before (Grantham 1974; Miyata, et al. 1979; Zhang 2000). However, we also consider other measures of physicochemical and evolutionary amino acid dissimilarity. We consider autosomal and X-linked loci separately since mutations on the X are hemizygous in males and there is some evidence that X-linked genes adapt faster (reviewed by Charlesworth et al. (Charlesworth, et al. 2018)).

We find that the rate of adaptive evolution relative to the mutation rate, ω_a, is significantly negatively correlated to both the difference in volume (Δ_{vol}) and polarity (Δ_{pol}), on both the autosomes and X-chromosome (Table 1; Figure 1A, B for autosomes; Figure S1A, B for X-chromosome) suggesting that the rate of adaptive evolution is higher between amino acids that are more physicochemically similar. The difference in volume and polarity are only weakly correlated (Spearman's $\rho = 0.13$, $p = 0.11$) and the two factors are independently correlated to ω_a in a multiple regression ($p<0.001$ for both factors on the autosomes and X).

There are many ways in which to measure the dissimilarity between amino acids, and there are over 500 dissimilarity matrices (Kawashima, et al. 2008). We find that ω_a is negatively correlated to the difference in amino acid properties in ~90% (476/531) of these matrices and significantly so in ~54% (286/531) matrices (Figure 2). ω_a is positively correlated to the difference in amino acid properties in 55 matrices but none of these correlations are significant.

So far, we have shown that the rate of adaptive evolution is higher between pairs of amino acids that are more similar in terms of volume and polarity. However, if dissimilar pairs of amino acids tend to be more common or have higher mutation rates, then the overall adaptive evolution might be dominated by substitutions of intermediate or large effect. As a consequence we calculated the total rate of adaptive substitution between each pair of amino acid as $\Omega_{a(ij)} = \omega_{a(ij)} \times (f_i + f_j) \times \mu_{ij}$ where $\omega_{a(ij)}$ is the ω_a between a pair of amino acids *i*

and j, f_i is the frequency of amino acid i and μ_{ij} is the mutation between them; we estimate the mutation rate from synonymous sites (e.g. if the amino acids are separated by a C<>T transition, we estimate the C<>T mutation rate from synonymous sites). If we plot the cumulative number of adaptive amino acids substitutions as a function of the difference in volume and polarity we find the relationship is concave suggesting that small substitutions dominate the adaptive process, when we take into account the frequencies and mutation rates of the amino acids (Figure 3A for autosomes; Figure S2A for the X-chromosome).

The fact that more similar amino acids have higher rates of adaptive evolution strongly suggests that the proportion of mutations that are adaptive is also higher amongst more similar amino acids, since more similar amino acids are likely to be subject to weaker positive selection and hence have lower fixation probabilities. We indeed observe this; λ_a is significantly negatively correlated to the difference in volume and polarity on both the X and autosomes (Table 1). If we calculate the overall rate of advantageous mutation for each pair of amino acids, taking into account the frequency of the amino acids and their mutation rate as $\Lambda_{a(ij)} = \lambda_{a(ij)} \times (f_i + f_j) \times \mu_{ij}$ and plot the cumulative, we again find that it is concave (Figure S3A, C).

Polarity and volume only explain some of the variance in ω_a and λ_a, particularly amongst amino acids that are similar in volume or polarity. This is not surprising; volume and polarity are just two measures of amino acid dissimilarity and there are many qualities that are difficult to quantify – for example the ability to form disulphide bridges. Alternative measures of amino acid dissimilarity are evolutionary measures such as the ratio of non-synonymous to synonymous polymorphisms (p_N/p_S) and the derived allele frequency of non-synonymous relative to synonymous polymorphisms (DAF_N/DAF_S). Both of these statistics are expected to be higher for amino acids that are more similar because they are expected to decline as the strength of selection against deleterious mutations increases. Consistent with this we find that p_N/p_S and DAF_N/DAF_S are negatively correlated to the difference in volume and polarity (Table 1; Figure 4).

Our two evolutionary measures of amino acid dissimilarity, p_N/p_S and DAF_N/DAF_S are not statistically independent of our measures of adaptive evolution, since polymorphism data is

used to estimate the rate of adaptive evolution; sampling error will therefore tend to induce correlations between ω_a, p_N/p_S and DAF_N/DAF_S. To overcome this, we resampled the SFS using a hypergeometric distribution to generate two SFSs, one of which was used to estimate p_N/p_S and DAF_N/DAF_S, and the other which was used to estimate the DFE and the rate of adaptive evolution. This procedure removes the non-independence due to sampling error, although we note that p_{N1}/p_{S1} and p_{N2}/p_{S2}, are very highly correlated to each other suggesting that there is relatively little sampling error relative to the systematic variance in p_N/p_S (autosome Spearman's $\rho = 0.96$, $p<0.001$; X-chromosome Spearman's $\rho = 0.86$, $p<0.001$); the correlation between DAF_{N1}/DAF_{S1} and DAF_{N2}/DAF_{S2} is also substantial on the autosomes (autosomes, Spearman's $\rho = 0.69$, $p<0.001$; X-chromosome Spearman's $\rho = 0.34$, $p = 0.003$). We find that ω_{a1} is significantly positively correlated to p_{N2}/p_{S2} and DAF_{N2}/DAF_{S2} (Table 1, Figure 1C, D). This is consistent with the pattern seen for volume and polarity; amino acids which are more similar in terms of the fitness effects, have high values of p_N/p_S and DAF_N/DAF_S, and higher rates of adaptive evolution. We also find that the proportion of mutations that are adaptive, λ_{a1}, is positively correlated to p_{N2}/p_{S2} and DAF_{N2}/DAF_{S2} (Table 1), again consistent with the pattern seen for polarity and volume. If we calculate the overall rates of adaptive substitution, $\Omega_{a(ij)}$, and mutation, $\Lambda_{a(ij)}$ and plot the cumulatives against the ranks of the p_{N2}/p_{S2} and DAF_{N2}/DAF_{S2} values in reverse order, we again observe concave functions (Figure 3B, S2B, S3B, D). Note that we plot the cumulatives against the rank, because p_N/p_S and DAF_N/DAF_S are not simple linear functions of the strength of selection, and we plot them in reverse order because large values correspond to more similar amino acids.

We have shown that both the rate of advantageous mutation and substitution is higher amongst amino acids that are more similar, where we have measured similarity both in terms of physicochemical and evolutionary differences. Finally, we would also like to know whether similar or dissimilar amino acids contribute more overall to adaptation. This question only makes sense phrased in terms of fitness. In principle, we can estimate the contribution of each amino acid pair to the change in fitness by multiplying the rate of adaptive evolution by the mean strength of selection acting on the advantageous substitutions. In principle it is possible to estimate the mean strength of selection from the

site frequency spectrum, with or without considering the rate of substitution (Schneider, et al. 2011; Tataru, et al. 2017; Tataru and Bataillon 2019). In practice, very large amounts of data are required. We find that our estimate of the strength of selection acting on advantageous mutations is uncorrelated to either the difference in volume, polarity, p_N/p_S or DAF_N/DAF_S (Table 1), which suggests that either the strength of selection acting on advantageous mutations is uncorrelated to the similarity of the amino acids, which seems unlikely, or that we cannot estimate the selection strength accurately enough. To assess the sampling error involved in estimating the strength of selection we bootstrapped the data 100 times for the 5 amino acid pairs for which we have the most non-synonymous polymorphisms. Despite having over 1500 non-synonymous polymorphisms in each case we find the confidence intervals span more than one order of magnitude (Figure S4). The reason for this uncertainty is evident upon a visual inspection of the SFSs (Figure S5). Under a model in which non-synonymous mutations are neutral or deleterious the ratio of the non-synonymous and synonymous SFS is expected to be a declining function. However, if there are advantageous mutations the ratio of SFS can be U-shaped and the uptick in the ratio at high allele frequencies contains information about the rate of advantageous mutation and the strength of selection acting upon those mutations (Schneider, et al. 2011; Tataru, et al. 2017). This signature is subtle and the ratio of the SFSs is too erratic to infer anything about the strength of selection acting on advantageous mutations (Figure S5).

Discussion

We have investigated whether the rate of advantageous mutation and substitution depends on the similarity of amino acids. We find that pairs of amino acids that are more similar have higher rates of advantageous mutation and substitution. The adaptive process therefore seems to be dominated by mutations and substitutions of small effect. This is true when we consider the amino acid pairs individually and when we take into account their frequencies and mutation rates. However, we have been unable to ascertain whether the overall change in fitness is dominated by small or large mutations. Using the analogy from the introduction, we have established that the supply of bricks is dominated by small bricks and that our builder's preferences are such that the wall contains more small than large bricks. However,

we have been unable to establish whether the wall is largely made of small or large bricks because we could not quantify the relative size of large vs small bricks.

Our work builds on the work of Grantham (Grantham 1974) and Miyata et al. (Miyata, et al. 1979) who showed, more than 40 years ago, that the rate of evolution is faster between amino acids that are more similar in their physicochemical properties. This might have been because more dissimilar amino acids have lower rates of adaptive evolution, lower rates of neutral evolution or both. We have shown that it is in part due to a lower rate of adaptive evolution (Table 1), but we can also test whether the rate of non-adaptive evolution $\omega_{na} = d_N/d_S - \omega_a$ (where d_N and d_S are rates of non-synonymous and synonymous divergence, respectively) (Galtier 2016) is correlated to amino acid dissimilarity. We find that ω_{na} is negatively correlated to the difference in volume or polarity, and positively correlated to p_N/p_S and DAF_N/DAF_S (Table 1). The fact that both the rate of adaptive and non-adaptive evolution decreases with increasing dissimilarity between amino acids suggests that the proportion of substitutions that are adaptive, α, might be relatively constant. We find, however, the proportion of substitutions that are adaptive, α, is significantly negatively correlated p_N/p_S and DAF_N/DAF_S and significantly positively correlated to the difference in polarity on the X-chromosome (Table 1); i.e. the proportion of substitutions that are adaptive is lower amongst amino acids that are more similar.

This latter result is consistent with the findings of Gojobori et al. (Gojobori, et al. 2007). They found that the fixation index, a statistic related to α, the proportion of non-synonymous substitutions fixed by positive selection, was negatively correlated to p_N/p_S in humans – i.e. the proportion of adaptive substitutions was higher amongst amino acids that were more dissimilar in evolutionary terms. However, it should be noted that they only took account of slightly deleterious mutations by removing rare variants, and that they found no evidence of adaptive evolution for most amino acid pairs.

We have used the method of Schneider et al. (Schneider, et al. 2011) to estimate the rate of adaptive evolution and its constituent parts. It is possible that some our results might be due to biases in the method, so to investigate we re-estimated the rate of adaptive

evolution using two alternative methods – the second method proposed by Eyre-Walker and Keightley (Eyre-Walker and Keightley 2009), which uses the method of Eyre-Walker et al. (Eyre-Walker, et al. 2006) to estimate the DFE, and the method of Tataru and Bataillon (Tataru and Bataillon 2019). The Schneider method estimates the DFE, the rate of adaptive mutation and the strength of selection acting upon the advantageous mutations using a combination of the SFS and the divergence between species, modelling demography explicitly using a three-epoch model. In contrast, the Eyre-Walker-Keightley method estimates the DFE and the rate of adaptive evolution using the SFS and the divergence between species modelling demography by a series of nuisance parameters. The method of Tataru et al. also models demography using nuisance parameters, but it estimates the DFE, the rate of adaptive mutation and the strength of selection using only the SFS. Hence, the three methods model demography in different ways and either do, or do not, use divergence data in their estimation of the rate of adaptive evolution.

As in the main analysis, we find that ω_a is negatively correlated to the difference in polarity and volume, and positively correlated to p_N/p_S and DAF_N/DAF_S using all methods (Tables S1 and S2); furthermore, we find the cumulative of Ω_a is concave (Figure S7 and S8). These results confirm that adaptive evolution is dominated by substitutions between amino acids that are relatively similar. We also find that ω_{na} is negatively correlated to the difference in polarity and volume, and positively correlated to p_N/p_S and DAF_N/DAF_S using all methods (Tables S1 and S2). However, we see differing patterns for α. The proportion of adaptive substitutions is significantly positively correlated to the difference in polarity and volume, and significantly negatively correlated to p_N/p_S and DAF_N/DAF_S using the Schneider and Tataru-Bataillon methods. However, using the Eyre-Walker-Keightley method we observe this pattern for the X-chromosome, but the opposite pattern for the autosomes. We also fail to find a significant correlation between the rate of adaptive mutation, λ_a, and any measure of the difference between amino acids, except a weakly significant negative correlation between λ_a and DAF_N/DAF_S on the autosomes when using the Tataru-Bataillon method, contrary to what we observe using the Schneider method (Table S1); the Eyre-Walker-Keightley method does not estimate λ_a. This might be because the Tataru-Bataillon method

only uses the SFS to infer the rate of advantageous mutation, and the SFS are subject to quite substantial levels of sampling error (see above).

Our results may explain the findings of Bazykin and Kondrashov (Bazykin and Kondrashov 2012) and Campos et al. (Campos, et al. 2017). Bazykin and Kondrashov (2012) observed that the rate of adaptive amino acid substitution was higher in regions of the gene that were less conserved. Campos et al. (Campos, et al. 2017) estimated that the rate of adaptive mutation was lower in more constrained genes, and surprisingly that the strength of selection acting upon those mutations was also weaker. Together these two inferences suggest that constrained genes would also undergo lower rates of adaptive substitution. Hence both analyses mirror at the gene and sub-gene level what we observe at the amino acid level. This raises the question whether genes and parts of genes adapt slowly because of the amino acids they contain, or whether certain amino acids have low rates of adaptive evolution because they tend to be found in genes and parts of genes that have low rates of adaptation. The fact that we observe strong correlations between rates of adaptive evolution and physicochemical properties suggests the former is at least partly true; genes and parts of genes that are constrained undergo low rates of adaptive evolution because they contain amino acids such as glycine which is small, leading to large volume differences, with amino acids that are one mutational step removed from it.

It is striking that much of the variance between amino acids in their rate of adaptive evolution can be explained in terms of p_N/p_S. Given that polymorphism data is expected to be dominated by neutral and slightly deleterious genetic variation, p_N/p_S is an estimate of the proportion of mutations that are effectively neutral and hence 1- p_N/p_S is a measure of the proportion of mutations that are deleterious. In part the correlation between ω_a and p_N/p_S is not surprising; as amino acids become more different so we expect the proportion of mutations that are effectively neutral to decline, and this is also likely to lead to a reduction in the proportion of mutations that are advantageous, as we have shown (Table 1). However, we might have also expected advantageous mutations between dissimilar amino acids to be more strongly selected (though see (Campos, et al. 2017)). We have been unable to ascertain whether this is the case ($N_e s_a$ is not significantly correlated to any measure of dissimilarity). However, we can conclude that the strength of selection acting

upon advantageous mutations either decreases as amino acid dissimilarity increases or stays constant, neither of which is very likely, or that it increases, but at a low rate, because the rate of adaptive substitution declines as amino acid similarity decreases; i.e. if the strength of selection acting upon advantageous mutations increased rapidly with increasing amino acid dissimilarity then the rate of adaptive evolution would be greater amongst more dissimilar amino acids, even though the proportion of mutations that are adaptive declines as amino acids become more dissimilar.

A potential problem in any analysis that uses the McDonald-Kreitman (MK) approach to estimate the rate of adaptive evolution are differences between the current N_e and the N_e during the divergence phase of evolution, if there is a class of mutations that are slightly deleterious (McDonald and Kreitman 1991; Eyre-Walker 2002). If the current N_e, which is relevant to the polymorphism data, is greater than the N_e for the divergence data then MK approaches will tend to overestimate the rate of adaptive evolution; the bias can be such that a signal of adaptive evolution can be detected even when there is no adaptive evolution occurring (McDonald and Kreitman 1991; Eyre-Walker 2002). It is not possible for us to rule out this as an explanation for the patterns we observe; the correlation between the rate of adaptive evolution and amino acid similarity might simply be a consequence of increasing population size. Similar analyses in other species are required.

The method that we have used to estimate the rate of adaptive evolution assumes that synonymous mutations are neutral, whereas selection is known to act upon synonymous codon use in some *Drosophila* species (Shields, et al. 1988; Akashi 1995). However, such selection is unlikely to affect our results because the rate of adaptive evolution is estimated using synonymous data that is common to multiple amino acid pairs that are separated by a particular type of mutation (e.g. C<>T). Selection on synonymous codon use could potentially affect the absolute rate of adaptive evolution but it's not expected to affect the pattern between pairs of amino acids. To investigate further we ran an analysis of covariance regressing ω_a against the difference in volume and polarity, with mutational type as a fixed effect (in effect fitting a series of parallel planes of ω_a against the difference in volume and polarity for each mutational type). We find that ω_a is significantly correlated to the difference in both volume ($p<0.001$) and polarity ($p<0.001$). It is also possible that biased

gene conversion could affect our results so we repeated the ANCOVA restricting our analysis to GC-conservative mutational types and again find that ω_a is significantly correlated to the difference in polarity ($p<0.001$) and volume ($p<0.001$).

Although we have shown that more similar amino acids undergo higher rates of advantageous mutation and substitution this does not directly address the underlying question of whether adaptive evolution is dominated by small or large effect mutations for two reasons. First, we have only considered amino acid mutations, but much adaptive evolution might proceed through regulatory changes (King and Wilson 1975; Andolfatto 2005). Second, underlying each amino acid pair is a distribution of effects; so, although we have shown that the average rate of advantageous mutation and substitution is correlated to measures of amino acid similarity, this does not imply that the underlying distribution, the distribution obtained by combining the distributions from each pair of amino acids, has the same shape. Overall, the adaptive process might be dominated by mutations and substitutions of intermediate effect, but the mean for each of the amino acid distributions is such that they lie to the right of mode of the underlying distribution (Figure S6).

In conclusion, whether evolution proceeds by large or small steps is a long-standing question. We have provided evidence that the adaptation of protein coding sequences is dominated by amino acid mutations that are of small effect.

Material and methods

Data and filtering

A population dataset of Zambian *D. melanogaster* sequences was taken from Lack et al. (Lack, et al. 2015). In total, the dataset consists of 197 sequences for each autosome and 196 sequences for the X chromosome. Sequences were annotated using the reference genome annotation of *D. melanogaster*} (r5.57 from http://www.flybase.org/) and subsequently masked for all non-coding regions to exclude genomic regions where coding and non-coding sequences overlap. Codon alignments were then extracted using a custom Python script. The alignment between the *D. melanogaster*, *D. simulans* and *D. yakuba* reference sequences was taken from Hu et al. (Hu, et al. 2013). Coding sequences which contained premature stop codons in the *D. melanogaster* reference sequence were excluded from the analysis.

Amino acid polarity scores and volumes were taken from the literature. Additionally, we analyzed other amino acid distance measures using data available in the AAindex1 database (Kawashima, et al. 2008). Specifically, for each index in the database, we calculated the physicochemical distance for all amino acid pairs under consideration, as the absolute difference. Indices which contained missing values for any amino acid were excluded from the analysis.

Parameter inference

We used three methods to estimate the rate of adaptive evolution for all 75 pairs of amino acids separated by a single mutational step: the method of Schneider et al. (Schneider, et al. 2011) using the software DFE-alpha version 2.16 (http://www.homepages.ed.ac.uk/pkeightl/software.html), the second method presented by Eyre-Walker and Keightley (Eyre-Walker and Keightley 2009) using the software DoFE version 3 (http://www.lifesci.susx.ac.uk/home/Adam_Eyre-Walker/Website/Software.html) and the polyDFE model B method of Tataru and Bataillon (Tataru and Bataillon 2019) (https://github.com/paula-tataru/polyDFE). The methods of Schneider et al. and Tataru and Bataillon require the unfolded site frequency spectrum (SFS) from a class of sites subject to selection, here non-synonymous sites, and a class of sites in which mutations are neutral, here synonymous sites. Inference of the unfolded site frequency spectrum for each of the

site classes was obtained by the method of Keightley et al. (Keightley, et al. 2016) (http://www.homepages.ed.ac.uk/pkeightl/software.html), using *D. simulans* and *D. yakuba* as outgroups for polarization of *D. melanogaster* sites into ancestral and derived allelic states. Although the dataset contains 197 and 196 lines for the autosomal and X-linked loci, we down-sampled the data to 20 lines. The subsampling step was necessary due to the limited size of the transition matrix used by the DFE-alpha program for estimating the demography parameters. Most amino acid pairs are separated by one of the six different mutational types. To estimate the rate of adaptive substitution we compared the SFS for a particular amino acid pair, say proline and threonine, which are separated by a C<>A change with synonymous data from 4-fold degenerate codons separated only by C<>A mutations ($SFS_{4F(C<>A)}$). For amino acids separated by more than one mutational type we calculated a weighted average SFS from the SFSs for the mutational types at 4-fold sites, weighting by the frequency of the respective codons. For example, leucine and valine are separated by C<>G and T<>G. The synonymous SFS used to estimate the rate of adaptive substitution was estimated as $SFS_{4F(weighted)}$ =(($f_{TTA} + f_{GTA} + f_{TTG} + f_{GTG}$) x $SFS_{4F(T<>G)}$ + ($f_{CTT} + f_{GTT} + f_{CTC} + f_{GTC}$…etc) x $SFS_{4F(C<>G)}$) / ($f_{TTA} + f_{GTA} + f_{TTG} + f_{GTG} + f_{CTT} + f_{GTT} + f_{CTC} + f_{GTC}$…etc) .

Six parameters were estimated using the method of Schneider et al. (Schneider, et al. 2011) for each of the 75 non-synonymous site classes: the proportion of adaptive substitutions α, the rate of adaptive evolution relative to the mutation rate, ω_a, the distribution of fitness effects for slightly deleterious mutations (DFE) modelled as a gamma distribution with the shape parameter β and the mean as the average selection strength against deleterious mutations, multiplied by the effective population size, $N_e\bar{s}_d$, the fitness effect of adaptive mutations, $N_e s_a$, as well as their proportion λ_a. The demography parameters necessary as input into the DFE-alpha program were inferred from the synonymous SFSs, assuming a 3-epoch model, as implemented in DFE-alpha. Similar selection parameters were estimated using the method of Tataru and Bataillon (Tataru and Bataillon 2019), but demography is modelled using nuisance parameters. The second method of Eyre-Walker and Keightley (Eyre-Walker and Keightley 2009) also uses nuisance parameters to model demography, but this method only estimates ω_a, β and $N_e\bar{s}_d$.

The average segregating frequency of polymorphisms for each site class was calculated as $DAF = (\sum_i iq_i)/\sum q_i$, where q_i represents the number of sites segregating at frequency i in the sample of sequences; $1 \leq i \leq 19$, as we construct the SFSs from 20 sequences.

Acknowledgements

JB was funded by the Austrian Science Fund (FWF, W1225-B20). We are grateful to referees for their comments.

Literature

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139:1067-1076.
Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149-1152.
Barrett RD, MacLean RC, Bell G. 2006. Mutations of intermediate effect are responsible for adaptation in evolving Pseudomonas fluorescens populations. Biol Lett 2:236-238.
Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. Trends Ecol Evol 23:38-44.
Bataillon T, Zhang T, Kassen R. 2011. Cost of adaptation and fitness effects of beneficial mutations in Pseudomonas fluorescens. Genetics 189:939-949.
Bazykin GA, Kondrashov AS. 2012. Major role of positive selection in the evolution of conservative segments of Drosophila proteins. Proc Biol Sci 279:3409-3417.
Bell G. 2009. The oligogenic view of adaptation. Cold Spring Harb Symp Quant Biol 74:139-144.
Campos JL, Zhao L, Charlesworth B. 2017. Estimating the parameters of background selection and selective sweeps in Drosophila in the presence of gene conversion. Proc Natl Acad Sci U S A 114:E4762-E4771.
Charlesworth B, Campos JL, Jackson BC. 2018. Faster-X evolution: Theory and evidence from Drosophila. Mol Ecol.
Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. Genetics 162:2017-2024.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26:2097-2108.
Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891-900.
Ferris MT, Joyce P, Burch CL. 2007. High frequency of mutations that expand the host range of an RNA virus. Genetics 176:1013-1022.
Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. PLoS Genet 12:e1005774.
Gojobori J, Tang H, Akey JM, Wu CI. 2007. Adaptive evolution in humans revealed by the negative correlation between polymorphism and fixation pahses of evolution. Proc. Natl. Acad. Sci. USA 104:3907-3912.
Gossmann T, Song B-H, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Fialtov DA, Eyre-Walker A. 2010. Genome wide analyses reveal little evidence of adaptive evolution in many plant species. Mol. Biol. Evol. 27:1822-1832.
Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science 185:862-864.
Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome Research 23:89-98.
Imhof M, Schlotterer C. 2001. Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. Proc. Natl. Acad. Sci. USA 98:1113-1117.

Kassen R, Bataillon T. 2006. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat Genet 38:484-488.
Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. 2008. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202-205.
Keightley PD, Campos JL, Booker TR, Charlesworth B. 2016. Inferring the Frequency Spectrum of Derived Variants to Quantify Adaptive Molecular Evolution in Protein-Coding Genes of Drosophila melanogaster. Genetics 203:975-+.
Kimura M. 1983. The neutral theory of molecular evolution. Cambridge, UK: Cambridge University Press.
King MC, Wilson AC. 1975. Evolution at 2 Levels in Humans and Chimpanzees. Science 188:107-116.
Kopp M, Hermisson J. 2009. The genetic basis of phenotypic adaptation II: the distribution of adaptive substitutions in the moving optimum model. Genetics 183:1453-1476.
Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, Langley CH, Pool JE. 2015. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics 199:1229-U1553.
MacLean RC, Buckling A. 2009. The distribution of fitness effects of beneficial mutations in Pseudomonas aeruginosa. PLoS Genet 5:e1000406.
Martin G, Lenormand T. 2008. The distribution of beneficial and fixed mutation fitness effects close to an optimum. Genetics 179:907-916.
Matuszewski S, Hermisson J, Kopp M. 2015. Catch Me if You Can: Adaptation from Standing Genetic Variation to a Moving Phenotypic Optimum. Genetics 200:1255-1274.
Matuszewski S, Hermisson J, Kopp M. 2014. Fisher's geometric model with a moving optimum. Evolution 68:2571-2588.
McDonald JH, Kreitman M. 1991. Adaptive evolution at the *Adh* locus in *Drosophila*. Nature 351:652-654.
McDonald MJ, Cooper TF, Beaumont HJ, Rainey PB. 2011. The distribution of fitness effects of new beneficial mutations in Pseudomonas fluorescens. Biol Lett 7:98-100.
Miyata T, Miyazawa S, Yasunaga T. 1979. Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219-236.
Orr HA. 2002. The population genetics of adaptation: the adaptation of DNA sequences. Evolution 56:1317-1330.
Orr HA. 1998. The Population Genetics of Adaptation: The Distribution of Factors Fixed during Adaptive Evolution. Evolution 52:935-949.
Perfeito L, Fernandes L, Mota C, Gordo I. 2007. Adaptive mutations in bacteria: high rate and small effects. Science 317:813-815.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:R208-215.
Rockman MV. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution 66:1-17.
Rokyta DR, Joyce P, Caudle SB, Wichman HA. 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat Genet 37:441-444.
Rozen DE, de Visser JA, Gerrish PJ. 2002. Fitness effects of fixed beneficial mutations in microbial populations. Curr Biol 12:1040-1045.
Sanjuan R, Moya A, Elena SF. 2004. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. Proc Natl Acad Sci U S A 101:8396-8401.

Schenk MF, Szendro IG, Krug J, de Visser JA. 2012. Quantifying the adaptive potential of an antibiotic resistance enzyme. PLoS Genet 8:e1002783.
Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427-1437.
Schoustra SE, Bataillon T, Gifford DR, Kassen R. 2009. The properties of adaptive walks in evolving populations of fungus. PLoS Biol 7:e1000250.
Seetharaman S, Jain K. 2014. Adaptive walks and distribution of beneficial fitness effects. Evolution 68:965-975.
Shields DC, Sharp PM, Higgins DG, Wright F. 1988. "Silent" sites in *Drosophila* are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5:704-716.
Tataru P, Bataillon T. 2019. polyDFEv2.0: Testing for invariance of the distribution of fitness effects within and across species. Bioinformatics.
Tataru P, Mollion M, Glemin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics 207:1103-1119.
Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J. Mol. Evol. 50:56-68.

		Δ_{vol}	Δ_{pol}	p_{N2}/p_{S2}	DAF_{N2}/DAF_{S2}
Autosome	p_N/p_S	-0.46***	-0.56***	0.96***	0.47***
	DAF_N/DAF_S	-0.38***	-0.41***	0.66***	0.69***
	β	0.29*	0.32**	-0.58***	-0.52***
	$N_e\bar{s}_d$	-0.20	-0.04	-0.08	-0.15
	α	0.05	-0.03	-0.31**	-0.25*
	ω_a	-0.47***	-0.53***	0.83***	0.54***
	ω_{na}	-0.50***	-0.54***	0.94***	0.51***
	λ_a	-0.35**	-0.37***	0.53***	0.54***
	$N_e s_a$	0.03	0.00	-0.04	-0.29*
X-chromosome	p_N/p_S	-0.37**	-0.50***	0.86***	0.44***
	DAF_N/DAF_S	-0.13	-0.20	0.40***	0.34**
	β	0.30**	0.20	-0.56***	-0.26*
	$N_e\bar{s}_d$	0.14	0.06	-0.28*	-0.11
	α	0.27*	0.20	-0.56***	-0.25*
	ω_a	-0.42***	-0.43***	0.79***	0.35**
	ω_{na}	-0.43***	-0.49***	0.88***	0.41***
	λ_a	-0.24*	-0.30**	0.38***	0.29*
	$N_e s_a$	-0.04	-0.05	0.16	-0.06

Table 1. Spearman rank correlation between estimates of rates of adaptive and non-adaptive evolution, the parameters of the DFE, and measures of amino acid dissimilarity. The symbols are as follows: p_N/p_S – the ratio of the number non-synonymous and synonymous polymorphisms per site; DAF_N/DAF_S – the ratio of the mean derived allele frequencies of non-synonymous and synonymous polymorphisms; β - the shape parameter of the DFE; $N_e\bar{s}_d$ – the mean strength of selection acting against deleterious mutations multiplied by the effective population size; α - the proportion of non-synonymous substitutions inferred to be advantageous; ω_a (ω_{na})– the rate of adaptive (non-adaptive) non-synonymous substitution rate relative to the mutation rate; λ_a – the proportion of

mutations inferred to be advantageous; $N_e s_a$ – the strength of selection inferred to be acting on advantageous mutations multiplied by the effective population size. To remove statistical non-independence between p_N/p_S and DAF_N/DAF_S and other variables we sampled the SFS to generate two independent SFSs. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

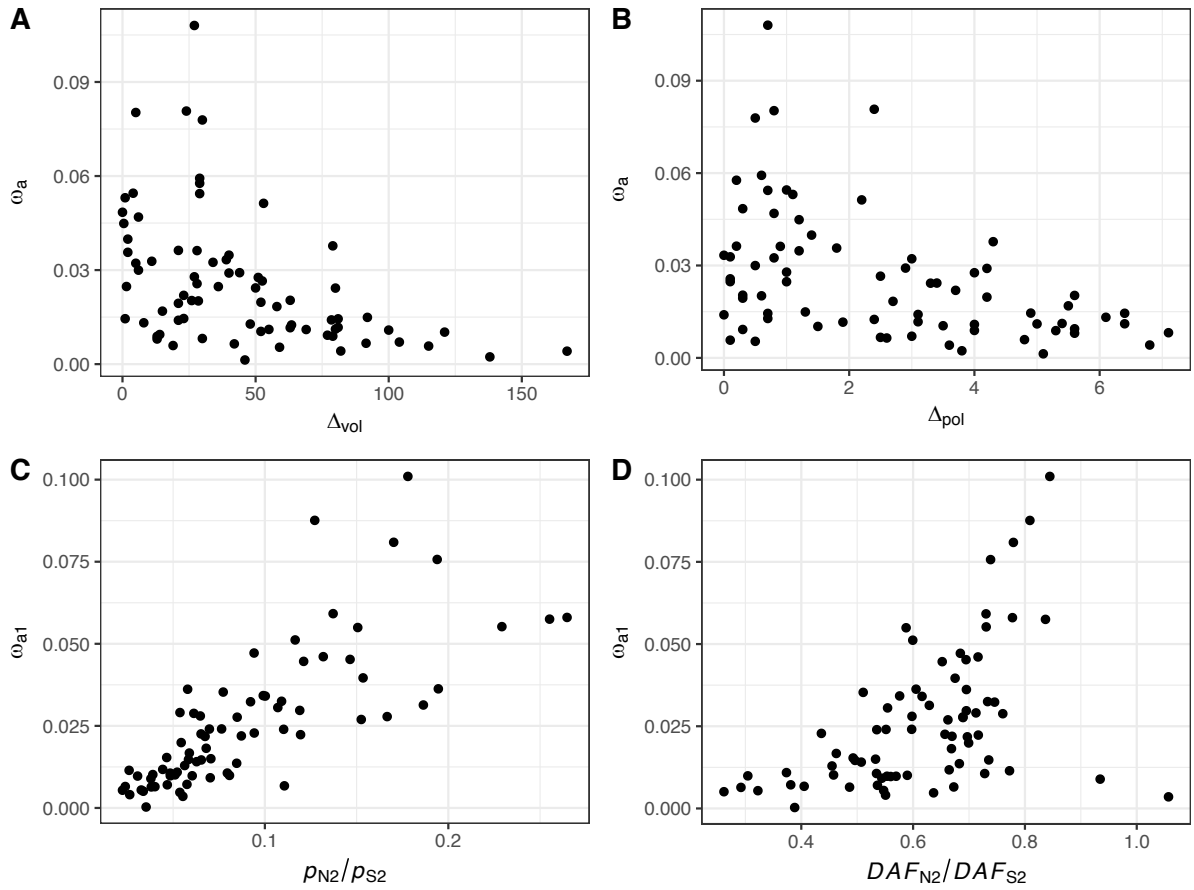

Figure 1. The autosomal rate of adaptive evolution relative to the mutation rate (ω_a) plotted against the difference in A) volume, B) polarity, C) the ratio of non-synonymous to synonymous polymorphisms, p_{N2}/p_{S2} and D) the ratio of the derived allele frequencies of non-synonymous and synonymous polymorphisms, DAF_{N2}/DAF_{S2}. In panels C) and D) the polymorphisms were split by sampling from a hypergeometric distribution; one set was used to calculate ω_{a1} the other the two polymorphism statistics.

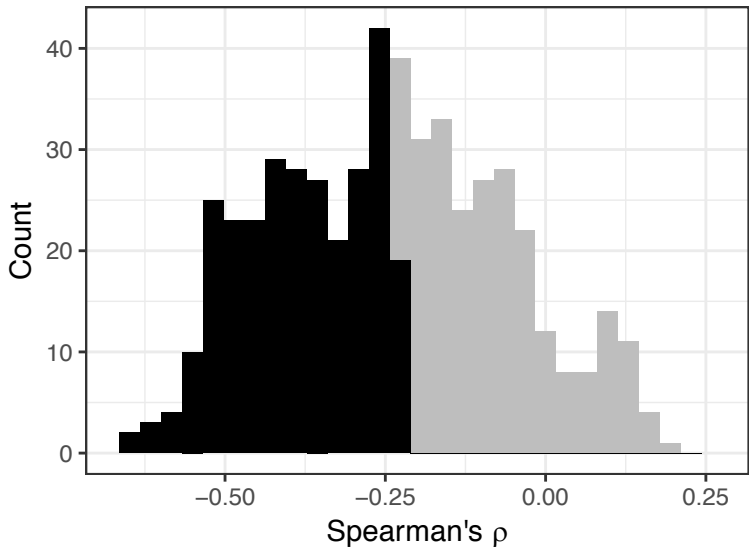

Figure 2. The distribution of Spearman rank correlations between ω_a and 531 amino acid dissimilarity matrices. The correlations in the darker shaded area are significant at 5%.

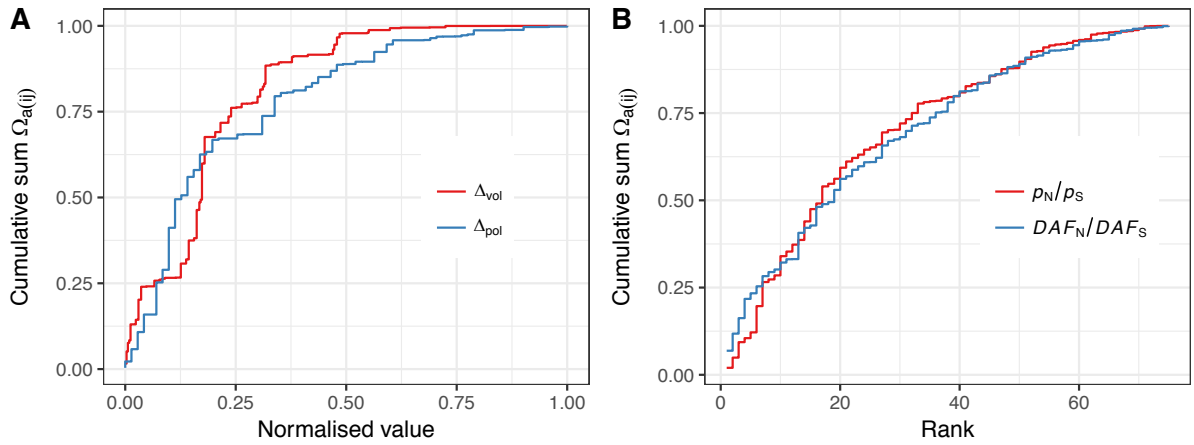

Figure 3. The cumulative number of adaptive substitutions on the autosomes contributed by each pair of amino acids versus A) the normalised difference in volume and polarity, and B) the reverse rank of p_N/p_S and DAF_N/DAF_S. The normalised difference in volume and polarity was calculated by subtracting the minimum difference, and then dividing by the maximum difference minus the minimum difference.

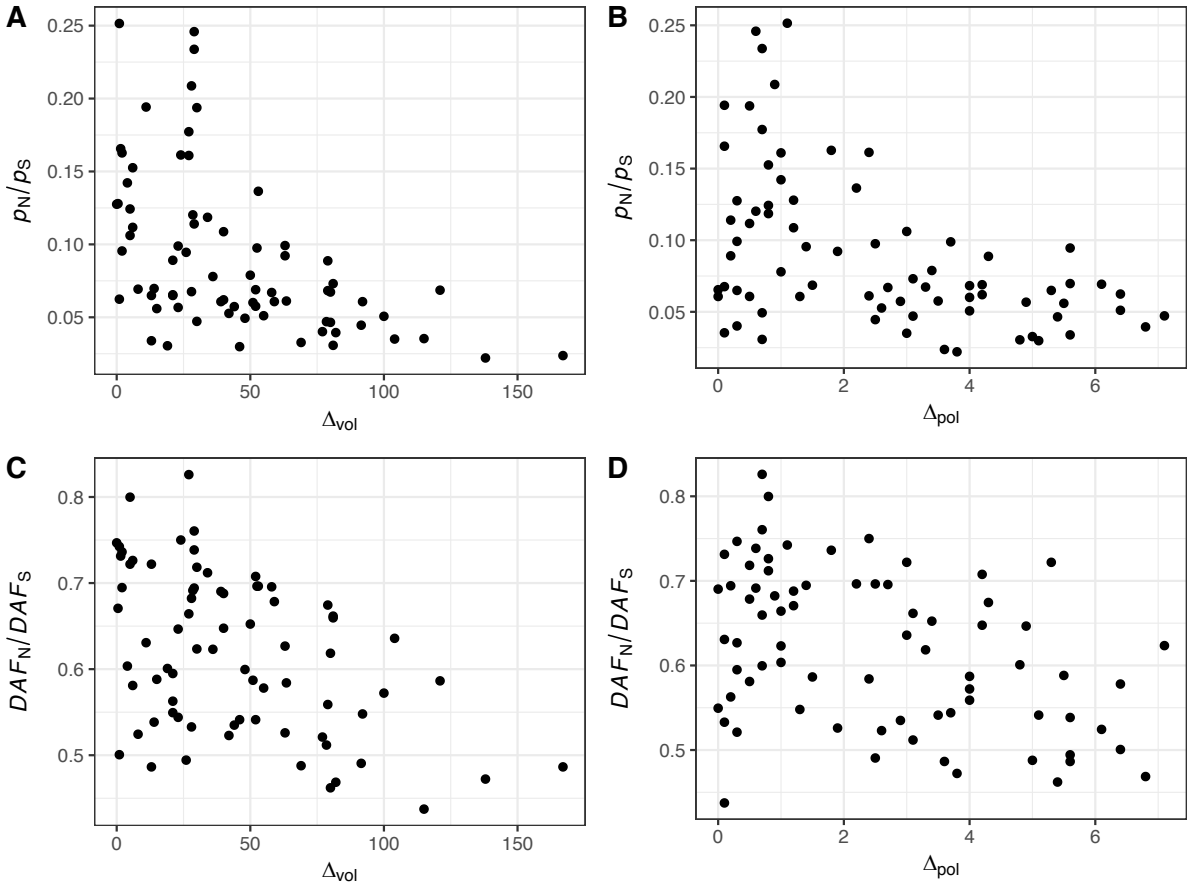

Figure 4. p_N/p_S and DAF_N/DAF_S plotted against the difference in volume and polarity for 75 pairs of amino acids for the autosomal data.